D352–D357 *Nucleic Acids Research, 2014, Vol. 42, Database issue* *Published online 5 December 2013*
doi:10.1093/nar/gkt1175

RepeatsDB: a database of tandem repeat protein structures

Tomás Di Domenico[1], Emilio Potenza[1], Ian Walsh[1], R. Gonzalo Parra[2], Manuel Giollo[1,3], Giovanni Minervini[1], Damiano Piovesan[1], Awais Ihsan[1,4], Carlo Ferrari[3], Andrey V. Kajava[5,6] and Silvio C.E. Tosatto[1,*]

[1]Department of Biomedical Sciences, University of Padua, 35131 Padova, Italy, [2]Department of Biological Chemistry, Universidad de Buenos Aires, Buenos Aires C1428EGA, Argentina, [3]Department of Information Engineering, University of Padua, 35121 Padova, Italy, [4]Department of Biosciences, COMSATS Institute of Information Technology, Sahiwal, Pakistan, [5]Centre de Recherches de Biochimie Macromoléculaire, CNRS, 34293 Montpellier Cedex 5, France and [6]Institut de Biologie Computationnelle, 34293 Montpellier Cedex 5, France

Received August 14, 2013; Revised October 28, 2013; Accepted October 29, 2013

ABSTRACT

RepeatsDB (http://repeatsdb.bio.unipd.it/) is a database of annotated tandem repeat protein structures. Tandem repeats pose a difficult problem for the analysis of protein structures, as the underlying sequence can be highly degenerate. Several repeat types haven been studied over the years, but their annotation was done in a case-by-case basis, thus making large-scale analysis difficult. We developed RepeatsDB to fill this gap. Using state-of-the-art repeat detection methods and manual curation, we systematically annotated the Protein Data Bank, predicting 10 745 repeat structures. In all, 2797 structures were classified according to a recently proposed classification schema, which was expanded to accommodate new findings. In addition, detailed annotations were performed in a subset of 321 proteins. These annotations feature information on start and end positions for the repeat regions and units. RepeatsDB is an ongoing effort to systematically classify and annotate structural protein repeats in a consistent way. It provides users with the possibility to access and download high-quality datasets either interactively or programmatically through web services.

INTRODUCTION

A large portion of proteins contain repetitive motifs, which are generated by internal duplications and frequently correspond to structural and functional units of proteins. Many repetitions in protein sequences can be identified by using different approaches (1–4). A more difficult problem for identification is however posed by repeats in protein structure, which can be highly degenerate (5,6). In fact, it is possible for a protein to maintain a repetitive structure even in the presence of massive amounts of point mutations (7). Several repeat families have been studied so far due to their relevance in different biological processes such as health (8), neurodevelopment (9) and protein engineering (10–12), to name just a few.

Repeats have been previously divided into five broad classes, primarily as a function of repeat length (13,14). At the lower end of the repeat length spectrum, i.e. less than five residues, very short repeats can either form insoluble aggregates (crystallites, class I) or long and winding helices of fibrous structures like collagen and α-helical coiled-coils (class II). At the other end of the spectrum, repeats containing $>\sim$50 residues appear to fold mostly as domains forming beads-on-a-string structures (class V). In between, for unit lengths of 5–40 residues, the known repeats can form either open elongated solenoids (class III) or closed toroids (class IV). Due to their fundamental functional importance, classes III and IV contain the most studied types of tandem repeat proteins. Solenoid folds appear to follow the distribution of repeat lengths rather closely, from all-beta (e.g. anti-freeze proteins) (15) to mixed alpha/beta (e.g. leucine-rich repeats) (16,17) to all-alpha structures (e.g. Armadillo and HEAT repeats) (18–20). They are characterized by some of the largest known autonomously folding domains, with 500 or more residues forming a single structure (21). Rapid addition or deletion of repeat units even between close homologs is of particular note for solenoid structures (22). Toroids on the other hand are restricted in overall size by their closed circular nature. Known toroid structures include the highly versatile TIM barrel and large outer membrane beta-barrels (23). Perhaps a more interesting fold is the beta propeller

*To whom correspondence should be addressed. Tel: +39 049 827 6269; Fax: +39 049 827 6260; Email: silvio.tosatto@unipd.it

(e.g. WD repeats), which can accommodate variable numbers of repeat units while maintaining a closed circular structure (24,25).

An open question regarding repeat proteins is the existence of other common structures that may have gone undetected. After all, the most common way to detect repeat families so far was to manually annotate the sequence family first and only afterwards visually recognize their structural repetitiveness. Such an approach is obviously difficult when dealing with the entire Protein Data Bank (PDB) (26), especially considering the many uncharacterized protein structures deposited by the main structural genomics consortia (27). The systematic description of repeat structures becomes a question of using automated methods to detect them in protein structures. This field is relatively new, with only few available methods. One of the first attempts was made by the Thornton group (28), but is unfortunately no longer available. Some methods (4,29–33) were developed to detect internal symmetries in proteins, but these may be difficult to adapt to the systematic classification of repeats. Recently, our group has developed RAPHAEL (34) in an attempt to fill the gap for repeat detection from structure. Widely used structural classifications such as CATH (35) and SCOP (36) also do not explicitly annotate repeats in protein structures, although it may be possible to leverage individual annotations to find similar repeats. Some databases exist for the detection of repeats from sequence (37–39), but usually these are limited to short tandem repeats and do not take into account divergent repeats, such as solenoids or toroids. The main domain sequence databases such as Pfam (40) and SMART (41) do not excel at the annotation of these repeat types either, as coverage is rather low and many repeat units go undetected. For Pfam most of the largest clusters of human sequence regions not covered were recently found to be repeats (42). To the best of our knowledge, no database or classification is currently available for repeat structures. This is the motivation for our present work, and we introduce RepeatsDB as a way to fill this gap. The database was developed to provide a central resource for the systematic annotation and classification of repeats. Given the fact that the structure-based search and classification of repeat proteins is more complete than on the basis of sequences or key words, our database will allow more accurate assignment of proteins with repeats to the corresponding families. For example, it will be used to suggest a better subdivision of alpha-solenoid proteins where at present the boundaries between the structures with Armadillo, HEAT, TPR and other repeat types are frequently blurred.

DATABASE DESCRIPTION

Data curation

The initial dataset for RepeatsDB was extracted from the PDB (43). Repeat candidates were identified from the reduced PDB dataset with RAPHAEL (34), which uses a geometric approach imitating the work of a human curator (score cutoff $\geq$1). The resulting dataset consisted of >10 000 repeat candidates, stored in the database as 'predicted' entries, which underwent a classification and curation process.

The dataset of predicted repeats was manually curated using a two-level annotation system. The first manual annotation level ('manually classified') classifies an entry into structural repeat class and subclass. This classification is based on previous work (14), where five classes of repeat structures are proposed, which are then further divided into subclasses. Class assignment is based mainly on repeat unit length and subclass assignment on secondary and tertiary structure features. The second manual annotation level ('detailed') consists in providing information about the start and end positions of the repeat units, repeat regions and/or insertions. We define a *repeat unit* as the smallest structural building block that is repeated to form a repeat region. A *repeat region* is a group of at least three repeat units. Inclusion of proteins with two repeat units would significantly complicate classification because many typical globular domains have this type of architecture. *Insertions* are non-repeated segments of structure that occur either inside a repeat unit or between two of them. These are particularly interesting because they break the repeat symmetry, and represent a challenge both for automatic detection and for the analysis of repeat structures (34).

Several curators annotated each protein undergoing manual classification by consensus. For first-level annotations, at least 75% of the curators had to agree in order for a protein to be included, otherwise it would be excluded and placed on a reserve list for future annotation. The rationale for this choice is that ambiguous cases are generally difficult to classify but may occasionally represent a novel repeat class. For second-level annotations, the threshold for consensus was at least 65% agreement (typically two of three curators). In case of discrepancy, an expert would arbitrate the final annotation based on the alternative proposals. Proteins with detailed annotations were also used to search for similar sequences in proteins from the PDB. Any PDB chain with at least 40% sequence identity and a coverage of at least 80% of the classified protein, belonging to the initial list of predicted entries, is added to the 'classified by similarity' annotation level. The similarity thresholds were selected to exclude possible false-positives (data not shown).

Implementation

RepeatsDB was designed with a multi-tier architecture, using separate modules for data management, data processing and presentation functions. To simplify development and maintenance, all tiers handle the common JSON (JavaScript Object Notation) format, thereby eliminating the need for data conversion. The MongoDB database engine is used for data storage and Node.js as middleware between data and presentation. RepeatsDB exposes its resources through RESTful web services, by using the Restify library for Node.js. The Angular.js framework and Bootstrap library were selected to provide the overall look-and-feel. Angular.js to Bootstrap integration is available through the angular-ui project. A customized

version of the BioJS (44) sequence component is used as sequence visualizer. Additional information is added to entries by querying the PDB web services at the structure and chain level. At the structure level, annotations like organism and experimental method used when resolving the structure are provided. At the chain level, secondary structure and links to other databases, among others. RepeatsDB offers users both graphical web interface access and RESTful web services from URL: http://repeatsdb.bio.unipd.it/

USING REPEATSDB

The user interface presents an intuitive tree-based browsing mechanism, where the root of the tree is the full database, second-level nodes repeat classes and third-level nodes subclasses. When clicking on a node, the user is presented with the list of RepeatsDB entries corresponding to the selected category. Each row of the list shows basic information about the entry, like its entry ID, title and organism. All annotated chains corresponding to an entry are displayed in a single page. The user interface presents a structure and sequence visualization widget (Figure 1). The user may choose to visualize the structure in four static images, or by using the 3D visualizer. If the entry features detailed annotations, the repeat regions, units and/or insertions are displayed using a combination of colours. The sequence visualization widget displays the sequence and secondary structure corresponding to the structure. It displays the same colour coding as the structure visualization widget, associating repeat annotations in the structure and sequence views. Additional

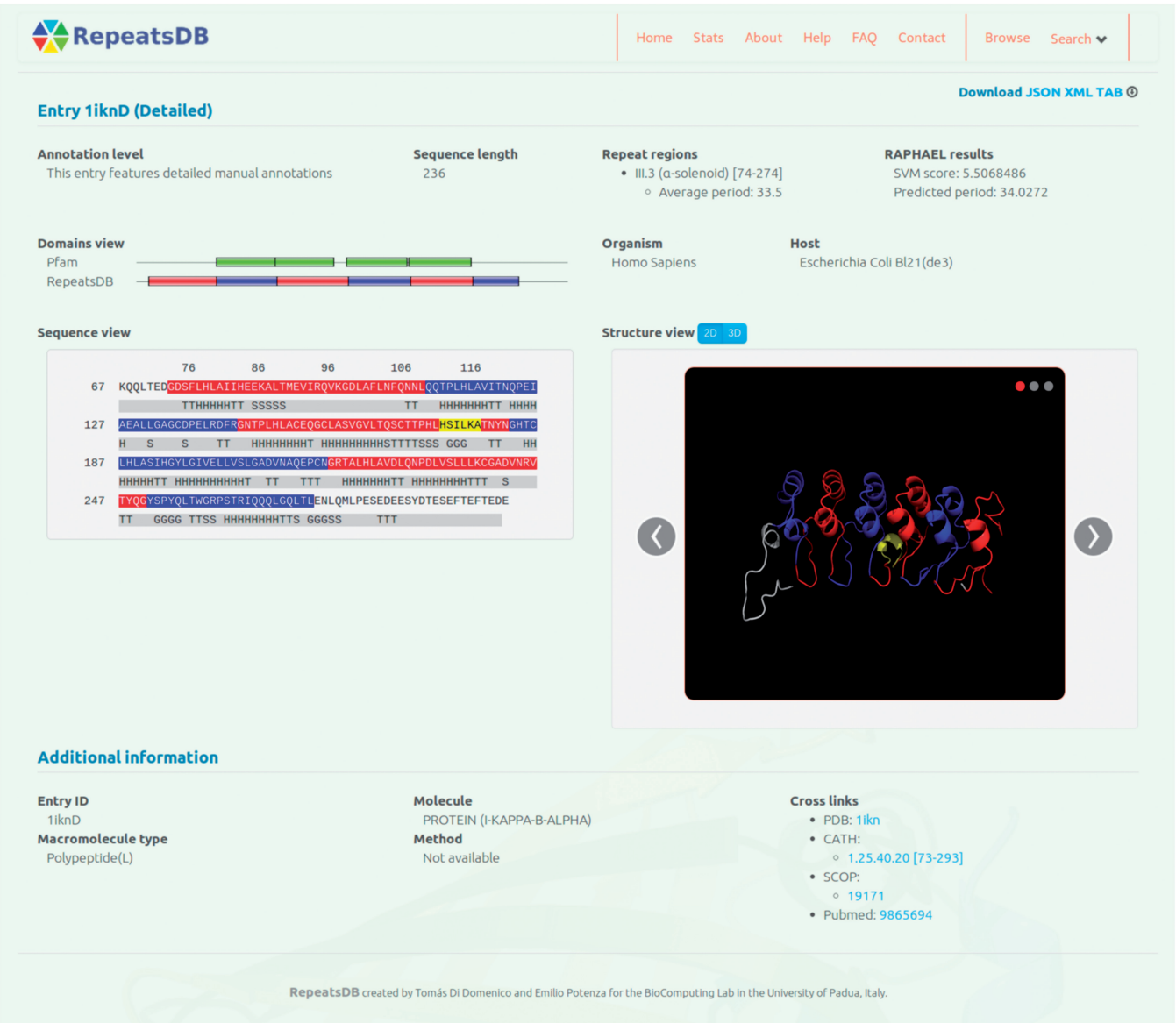

Figure 1. Screenshot of a sample RepeatsDB entry results page (PDB entry 1ikn). The sequence viewer and the structure viewer are shown in the middle of the page, towards the left and the right, respectively. Additional annotations at the structure and chain level are displayed, including links to other databases (above) and classifications (below).

information at the structure and chain levels is also provided.

The RepeatsDB search toolbar, available on top of every page, allows to search for entries either by database IDs or UniProt text query. The database ID search allows comma-separated PDB or UniProt IDs. The UniProt text search query uses the full UniProt search engine, see online documentation. RESTful web services are directly accessible through HTTP URLs. All data available on RepeatsDB are also available for programmatic access. Please refer to the 'Help' section of the website for details on using the RepeatsDB web services. Datasets can be downloaded in JSON, XML or text format using the browse function or RESTful web services.

Statistics

Analysis of the full PDB dataset yielded 10 745 repeats predicted by RAPHAEL, of which 2797 were finally classified into the RepeatsDB schema. Table 1 shows the distribution between classes and subclasses. The bulk of the annotations (~90%) consist of entries belonging to classes III and IV. No effort was made to balance the distribution of entries between classes in this initial release. As coverage increases in the future, we expect the balance to approximate the real distribution more closely, although it may be necessary to fine-tune RAPHAEL. Of the classified entries, 321 representatives of the entire dataset were annotated in detail with information about the start and end of repeat regions, repeat units and/or insertions (Table 1). It is interesting to note the different distribution of insertions between classes. Apparently, some classes such as β-solenoid (class III.1) or TIM barrels (class IV.1) have stronger propensity to accommodate insertions.

CONCLUSIONS AND FUTURE WORK

RepeatsDB's goal is to provide the community with a resource for high-quality tandem repeat protein structure annotations. The user can either interactively analyse his proteins of interest via the user interface, or create and download datasets for offline use. Far from being a static classification process, the annotation effort for the initial RepeatsDB dataset alone already motivated the extension of the original classification schema (14). Some of the curated structures, while clearly representing structural repeats, did not belong to any of the pre-defined subclasses. To allow them to be classified, subclasses IV.5 (α/β prism) and IV.6 (α-barrel) were added to the initial schema (14). Class V also underwent a re-classification according to the secondary structure content of the single domain repeats ('beads') to allow a broader classification range beyond individual repeat families, as the list of possible beads-on-a-string folds may be considerably larger than currently appreciated. The 'other' subclass was also added to allow collection of repeats that do not fit into the current classification scheme. RepeatsDB provides the community with a previously unavailable opportunity to easily create datasets of tandem repeat proteins. The detailed annotation subset further presents a unique opportunity to better understand the nature of tandem repeat proteins.

Beyond its initial release, RepeatsDB is a continuous effort to expand, revise and improve tandem protein repeat annotations. Predictions for new PDB structures are simple and fully automated, allowing regular database updates every 3 months. Manual curation of new entries for inclusion is also ongoing, aiming at regular and steady updates. Options to involve the community into the annotation process through crowdsourcing tools are currently being analysed. A main goal

Table 1. Statistics for RepeatsDB

Subclass	Name	Detailed	Classified (manually)	Classified (by similarity)	Predicted
I.1	Poly-alanine β structure	0	0	0	0
II.1	Collagen triple-helix	0	5	0	0
II.2	α helical coiled coil	23	38	69	0
III.1	β-solenoid	43	113	21	0
III.2	α/β solenoid	21	43	27	0
III.3	α-solenoid	48	246	631	0
III.4	Trimer of β spirals	7	0	13	0
III.5	Single layer anti-parallel β	4	3	0	0
IV.1	TIM-barrel	84	118	626	0
IV.2	β-barrel	8	1	8	0
IV.3	β-trefoil	20	0	29	0
IV.4	β-propeller	40	182	227	0
IV.5	α/β prism	0	17	0	0
IV.6	α-barrel	6	0	0	0
V.1	α-beads	2	1	0	0
V.2	β-beads	29	12	71	0
V.3	α/β-beads	3	3	1	0
V.other	Unknown subclass	3	0	4	0
UA	Unassigned	0	0	0	7948
	Total	321	749	1727	7948

The subclass name is shown together with the number of entries on each of the four annotation levels. Note that 'Unassigned' entries are automatically predicted by RAPHAEL and therefore not assigned to a specific class.

for future versions is the extension of the annotation of repeats at the sequence level, starting from annotation for intrinsically disordered regions from MobiDB (45). We anticipate that RepeatsDB should prove valuable towards the understanding of the sequence–structure relationship in tandem repeat proteins and their evolutionary relationship.

ACKNOWLEDGEMENTS

The authors are grateful to Diego Ferreiro for insightful discussions and to Manfred Sippl for his software tools.

FUNDING

FIRB Futuro in Ricerca [RBFR08ZSXY] and AIRC [MFAG 12740] (to S.T.); AIRC research fellow (to G.M.) and CONICET PhD student (to R.G.P.); Ministry of Education and Science of the Russian Federation, project 8831 (to A.V.K.) (in part). Funding for open access charge: FIRB Futuro in Ricerca [RBFR08ZSXY].

Conflict of interest statement. None declared.

REFERENCES

1. Wootton,J.C. (1994) Non-globular domains in protein sequences: automated segmentation using complexity measures. *Comput. Chem.*, **18**, 269–285.
2. Jorda,J. and Kajava,A.V. (2010) Protein homorepeats sequences, structures, evolution, and functions. *Adv. Protein Chem. Struct. Biol.*, **79**, 59–88.
3. Gribskov,M., McLachlan,A.D. and Eisenberg,D. (1987) Profile analysis: detection of distantly related proteins. *Proc. Natl Acad. Sci. USA*, **84**, 4355–4358.
4. Biegert,A. and Söding,J. (2008) *De novo* identification of highly diverged protein repeats by probabilistic consistency. *Bioinformatics*, **24**, 807–814.
5. Schaper,E., Kajava,A.V., Hauser,A. and Anisimova,M. (2012) Repeat or not repeat?—statistical validation of tandem repeat prediction in genomic sequences. *Nucleic Acids Res.*, **40**, 10005–10017.
6. Buard,J. and Vergnaud,G. (1994) Complex recombination events at the hypermutable minisatellite CEB1 (D2S90). *EMBO J.*, **13**, 3203–3210.
7. Andrade,M.A., Perez-Iratxeta,C. and Ponting,C.P. (2001) Protein repeats: structures, functions, and evolution. *J. Struct. Biol.*, **134**, 117–131.
8. Kajava,A.V. and Steven,A.C. (2006) Beta-rolls, beta-helices, and other beta-solenoid proteins. *Adv. Protein Chem.*, **73**, 55–96.
9. De Wit,J., Hong,W., Luo,L. and Ghosh,A. (2011) Role of leucine-rich repeat proteins in the development and function of neural circuits. *Annu. Rev. Cell Dev. Biol.*, **27**, 697–729.
10. Main,E.R., Lowe,A.R., Mochrie,S.G., Jackson,S.E. and Regan,L. (2005) A recurring theme in protein engineering: the design, stability and folding of repeat proteins. *Curr. Opin. Struct. Biol.*, **15**, 464–471.
11. Stefan,N., Martin-Killias,P., Wyss-Stoeckle,S., Honegger,A., Zangemeister-Wittke,U. and Plückthun,A. (2011) DARPins recognizing the tumor-associated antigen EpCAM selected by phage and ribosome display and engineered for multivalency. *J. Mol. Biol.*, **413**, 826–843.
12. Javadi,Y. and Itzhaki,L.S. (2013) Tandem-repeat proteins: regularity plus modularity equals design-ability. *Curr. Opin. Struct. Biol.*, **23**, 622–631.
13. Marcotte,E.M., Pellegrini,M., Yeates,T.O. and Eisenberg,D. (1999) A census of protein repeats. *J. Mol. Biol.*, **293**, 151–160.
14. Kajava,A.V. (2012) Tandem repeats in proteins: from sequence to structure. *J. Struct. Biol.*, **179**, 279–288.
15. Bateman,A., Murzin,A.G. and Teichmann,S.A. (1998) Structure and distribution of pentapeptide repeats in bacteria. *Protein Sci.*, **7**, 1477–1480.
16. Bella,J., Hindle,K.L., McEwan,P.A. and Lovell,S.C. (2008) The leucine-rich repeat structure. *Cell. Mol. Life Sci.*, **65**, 2307–2333.
17. Kobe,B. and Kajava,A.V. (2001) The leucine-rich repeat as a protein recognition motif. *Curr. Opin. Struct. Biol.*, **11**, 725–732.
18. Tewari,R., Bailes,E., Bunting,K.A. and Coates,J.C. (2010) Armadillo-repeat protein functions: questions for little creatures. *Trends Cell Biol.*, **20**, 470–481.
19. Kajava,A.V., Gorbea,C., Ortega,J., Rechsteiner,M. and Steven,A.C. (2004) New HEAT-like repeat motifs in proteins regulating proteasome structure and function. *J. Struct. Biol.*, **146**, 425–430.
20. Andrade,M.A., Petosa,C., O'Donoghue,S.I., Müller,C.W. and Bork,P. (2001) Comparison of ARM and HEAT protein repeats. *J. Mol. Biol.*, **309**, 1–18.
21. Kobe,B. and Kajava,A.V. (2000) When protein folding is simplified to protein coiling: the continuum of solenoid protein structures. *Trends Biochem. Sci.*, **25**, 509–515.
22. Björklund,A.K., Ekman,D. and Elofsson,A. (2006) Expansion of protein domain repeats. *PLoS Comput. Biol.*, **2**, e114.
23. Remmert,M., Biegert,A., Linke,D., Lupas,A.N. and Söding,J. (2010) Evolution of outer membrane beta-barrels from an ancestral beta beta hairpin. *Mol. Biol. Evol.*, **27**, 1348–1358.
24. Jawad,Z. and Paoli,M. (2002) Novel sequences propel familiar folds. *Structure*, **10**, 447–454.
25. Chaudhuri,I., Söding,J. and Lupas,A.N. (2008) Evolution of the beta-propeller fold. *Proteins*, **71**, 795–803.
26. Berman,H.M., Kleywegt,G.J., Nakamura,H. and Markley,J.L. (2013) The future of the protein data bank. *Biopolymers*, **99**, 218–222.
27. Dessailly,B.H., Nair,R., Jaroszewski,L., Fajardo,J.E., Kouranov,A., Lee,D., Fiser,A., Godzik,A., Rost,B. and Orengo,C. (2009) PSI-2: structural genomics to cover protein domain family space. *Structure*, **17**, 869–881.
28. Murray,K.B., Taylor,W.R. and Thornton,J.M. (2004) Toward the detection and validation of repeats in protein structure. *Proteins*, **57**, 365–380.
29. Parra,R.G., Espada,R., Sánchez,I.E., Sippl,M.J. and Ferreiro,D.U. (2013) Detecting repetitions and periodicities in proteins by tiling the structural space. *J. Phys. Chem. B.*, **117**, 12887–12897.
30. Marsella,L., Sirocco,F., Trovato,A., Seno,F. and Tosatto,S.C. (2009) REPETITA: detection and discrimination of the periodicity of protein solenoid repeats by discrete Fourier transform. *Bioinformatics*, **25**, i289–i295.
31. Szklarczyk,R. and Heringa,J. (2004) Tracking repeats using significance and transitivity. *Bioinformatics*, **20(Suppl. 1)**, i311–i317.
32. Heger,A. and Holm,L. (2000) Rapid automatic detection and alignment of repeats in protein sequences. *Proteins*, **41**, 224–237.
33. Abraham,A.L., Rocha,E.P. and Pothier,J. (2008) Swelfe: a detector of internal repeats in sequences and structures. *Bioinformatics*, **24**, 1536–1537.
34. Walsh,I., Sirocco,F.G., Minervini,G., Di Domenico,T., Ferrari,C. and Tosatto,S.C. (2012) RAPHAEL: recognition, periodicity and insertion assignment of solenoid protein structures. *Bioinformatics*, **28**, 3257–3264.
35. Sillitoe,I., Cuff,A.L., Dessailly,B.H., Dawson,N.L., Furnham,N., Lee,D., Lees,J.G., Lewis,T.E., Studer,R.A., Rentzsch,R. *et al.* (2013) New functional families (FunFams) in CATH to improve the mapping of conserved functional sites to 3D structures. *Nucleic Acids Res.*, **41**, D490–D498.
36. Andreeva,A., Howorth,D., Chandonia,J.M., Brenner,S.E., Hubbard,T.J., Chothia,C. and Murzin,A.G. (2008) Data growth and its impact on the SCOP database: new developments. *Nucleic Acids Res.*, **36**, D419–D425.
37. Jorda,J., Baudrand,T. and Kajava,A.V. (2012) PRDB: Protein Repeat DataBase. *Proteomics*, **12**, 1333–1336.
38. Luo,H., Lin,K., David,A., Nijveen,H. and Leunissen,J.A. (2011) ProRepeat: an integrated repository for studying amino acid tandem repeats in proteins. *Nucleic Acids Res.*, **40**, D394–D399.

39. Robertson,A.L., Bate,M.A., Androulakis,S.G., Bottomley,S.P. and Buckle,A.M. (2011) PolyQ: a database describing the sequence and domain context of polyglutamine repeats in proteins. *Nucleic Acids Res.*, **39**, D272–D276.
40. Punta,M., Coggill,P.C., Eberhardt,R.Y., Mistry,J., Tate,J., Boursnell,C., Pang,N., Forslund,K., Ceric,G., Clements,J. *et al.* (2012) The Pfam protein families database. *Nucleic Acids Res.*, **40**, D290–D301.
41. Letunic,I., Doerks,T. and Bork,P. (2011) SMART 7: recent updates to the protein domain annotation resource. *Nucleic Acids Res.*, **40**, D302–D305.
42. Mistry,J., Coggill,P., Eberhardt,R.Y., Deiana,A., Giansanti,A., Finn,R.D., Bateman,A. and Punta,M. (2013) The challenge of increasing Pfam coverage of the human proteome. *Database*, **2013**, bat023.
43. Rose,P.W., Bi,C., Bluhm,W.F., Christie,C.H., Dimitropoulos,D., Dutta,S., Green,R.K., Goodsell,D.S., Prlic,A., Quesada,M. *et al.* (2013) The RCSB Protein Data Bank: new resources for research and education. *Nucleic Acids Res.*, **41**, D475–D482.
44. Gómez,J., García,L.J., Salazar,G.A., Villaveces,J., Gore,S., García,A., Martín,M.J., Launay,G., Alcántara,R., del-Toro,N. *et al.* (2013) BioJS: an open source JavaScript framework for biological data visualization. *Bioinformatics*, **29**, 1103–1104.
45. Di Domenico,T., Walsh,I., Martin,A.J. and Tosatto,S.C. (2012) MobiDB: a comprehensive database of intrinsic protein disorder annotations. *Bioinformatics*, **28**, 2080–2081.